Acta Cryst. (1997). A**53**, 225–235

A Discrete Algorithm for Crystal Structure Prediction of Organic Molecules

DETLEF W. M HOFMANN[a]* AND THOMAS LENGAUER[a,b]

[a]*Institute for Algorithms and Scientific Computing, German National Research Center for Information Technology (GMD–SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany, and* [b]*Department of Computer Science, University of Bonn, Römerstrasse 164, 53117 Bonn, Germany. E-mail: detlef.hofmann@gmd.de*

(Received 3 April 1996; accepted 20 November 1996)

Abstract

A new algorithm, *FlexCryst*, is presented for fast crystal structure prediction. The algorithm differs from existing algorithms in that it performs the analysis on the basis of only a single molecule and uses potentials for scoring energy that are derived statistically from a set of data on molecular structures. In a first step, the algorithm creates various potential unit cells. In the second step, a set of candidates for translation vectors for corresponding crystals is generated. In the third step, the algorithm selects triples of candidate vectors to form potential crystal structures. The fourth step ranks the crystal structures with respect to their energy as estimated by a suitable scoring function. In the last step, the crystal structures are clustered according to a newly defined measure of similarity for crystal structures. At the moment, the program can handle only triclinic crystals with one molecule per asymmetric unit. The algorithm was tested on a set of 131 experimentally resolved crystals of space group $P1$ and 95 crystals of space group $P\bar{1}$ from the Cambridge Structural Database. For $P1$, in 129 cases (98%), the observed crystal structure is among the crystal structures generated by the algorithm. The run time of the algorithm is a few seconds per molecule on a standard workstation. For $P\bar{1}$, the experimental structure has been found among the proposed structures in 81 cases (85%). Owing to the more complex unit cell for this space group, the run time increases to about 2 h per molecule.

1. Introduction

The prediction of crystals formed by organic molecules is of interest for various reasons. With knowledge of the crystal structure, we can estimate the color and conductivity and derive from the space group nonlinear optical effects (Williams, 1983), pyroelectricity, ferroelectricity and piezoelectricity (Borchardt-Ott, 1976; Hahn & Klapper, 1995), ferromagnetism (Veciana, Cirujeda, Rovira & Vidal-Goncedo, 1995) and triboluminescence (Zink, 1978). An important advantage of using organic compounds is that their molecular structures may be tailored to enhance their desired physical properties.

Existing algorithms for crystal structure prediction are based on the atom–atom potentials approach. 50 years ago, this idea was formulated for organic molecules (Westheimer & Mayer, 1946) and applied to various organic reactions (Hill, 1948). It was transferred to crystal structures by Kitaigorodskii (1951). This approach is based on the assumption that the total interaction energy can be expressed as a sum of pairwise contributions between atoms of the molecules making up the crystal. In general, the potential contains three terms, which can be interpreted as a Coulombic term, the dispersion energy and the exchange repulsion energy. The second and the third terms together are called the *Lennard-Jones potential* or the *Buckingham potential*, depending on their analytical expression. Several combinations of these potentials are used in the literature: Buckingham potential alone (Filippini & Gavezotti, 1993), Buckingham potential and Coulombic term (Xiao & Williams, 1993) or Lennard-Jones potential and Coulombic term (Shoda, Yamahara, Okazaki & Williams, 1994). This approach towards estimating energy is tantamount to the existing force-field methods if we omit the terms for bonds, angles and dihedral angles. The parameters for the potentials are derived with *ab initio* methods or from experimental data. The heuristic form of the potentials requires parameter sets that are specific to restricted classes of organic molecules. After determination of the parameter set, the conformational space is searched for the global minimum. This is done by starting with the molecule, constructing the crystal in a certain space group and optimizing the degrees of freedom by different gradient methods, such as pseudo-annealing (Shoda, Yamahara, Okazaki & Williams, 1995), Newton–Raphson or steepest descent (Tajima *et al.*, 1995). Usually, the optimization starts with a random structure and searches a nearby minimum. This procedure is repeated several times, resulting in a set of local minima. These minima are then ranked with respect to their energy, as estimated by the potential, and the highest ranking, *i.e.* lowest-energy minimum, is taken to be the global minimum. This global minimum should coincide with the experimentally observed crystal structure. In practice, this can be a very difficult task, first, because the potentials are inaccurate and, second, because thousands of quite different local minima can fall within a very narrow energy range (40 kJ mole^{-1}), as has been witnessed for monosaccharides (van Eijck,

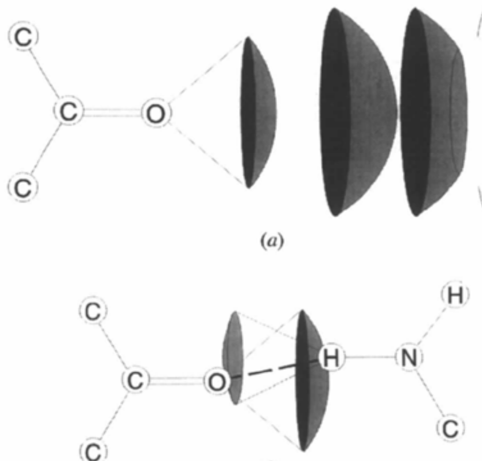
Acta Crystallographica Section A
ISSN 0108-7673 © 1997

Mooij & Kroon, 1995). Thus, today, crystal predictors are satisfied if the observed crystal structure is among the highest-ranking solutions found by the algorithm. A comprehensive overview of methods for predicting crystal structures is given by Desiraju (1989).

2. The algorithm *FlexCryst*

Our new algorithm does not use a search for local minima based on classical potentials in order to optimize conformations. Rather, we model the conformational space of the potential crystal structures discretely and perform a combinatorial search on this space. Our scoring function is a pair potential that is derived statistically from a subset of observed crystal structures taken from the Cambridge Structural Database (Allen & Kennard, 1993).

In the first step, the algorithm analyzes a given fixed conformation of the organic molecule using the approach taken by the program *FlexX* (Rarey, Kramer & Lengauer, 1995; Rarey, Kramer, Lengauer & Klebe, 1996; Rarey, Wefing & Lengauer, 1996). This program was developed for the prediction of the molecular interaction between protein receptors and small organic ligands. The approach models intermolecular interactions both geometrically and chemically. In general, an interaction is formed if specific geometric constraints are met by the interaction partners, which are functional groups. For each interaction formed, its contribution to energy is rated by a statistically derived scoring function. *FlexX* models hydrogen bonds, phenyl ring–phenyl ring, phenyl ring–methyl and phenyl ring–amide group interactions and is based on the work of Klebe & Mietzner (1994) and Böhm (1992, 1994). The geometric constraints have the same structure for all types of interaction. Specifically, an interaction between two molecular groups is represented by an *interaction center* – which is a point in space usually coinciding with the location of an atom or the center of a ring – and parts of a spherical *interaction surface* at a specific *interaction distance* around the interaction center, see Fig. 1. An interaction between two groups is formed if the interaction center of the first group lies on the interaction surface of the second group and *vice versa*. In order to keep the model discrete, we represent the interaction surface of the interaction partners by a finite point set. Based on this model, the algorithm generates dimers and ranks their stability. In general, there are six degrees of freedom for docking two rigid molecules: three for the translation and three for the rotation of the two molecules relative to each other.



(*a*)

(*b*)

Fig. 1. Modeling molecular interactions with *FlexX*. (*a*) Different shapes of interaction surfaces. (*b*) Model of a hydrogen bond.

In the following, we restrict ourselves to triclinic crystals with one molecule per asymmetric cell. Here we can distinguish between two cases, space groups $P1$ and $P\bar{1}$. For $P1$, the molecule is identical to the unit cell (without the lattice information) but in the case of $P\bar{1}$ the unit cell contains two molecules. These molecules are symmetry related by an inversion center. Therefore, in the case of $P\bar{1}$, we first search for possible inversion centers that map interaction centers of the molecule onto interaction surface points. A vector of one of these inversion centers can be calculated as the sum of one vector of an interaction surface point and a vector of an interaction center divided by two. Each of these unit cells is assigned a rank by our scoring function. The 50 highest-ranking cells are retained for further computation. After the selection of these unit cells, the interaction points are inverted as well. Each of the unit cells is defined by the molecule and one three-dimensional vector for the inversion center, in the case of $P\bar{1}$, and by the molecule only, in the case of $P1$.

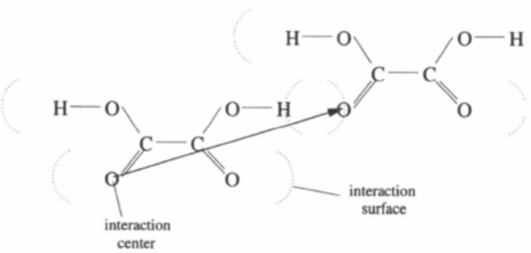


Fig. 2. An interaction between two unit cells, each containing one oxalic acid. The interaction center of one molecule lies on the interaction surface of the other molecule. The resulting translation vector is identical to the vector connecting the interaction center with the interaction surface inside the same molecule.

For a given unit cell, the remaining problem is to determine the three translation vectors spanning the crystal. In order to calculate possible translation vectors, the interaction center of one unit cell can be matched with a point on the interaction surface of a neighboring unit cell. This is tantamount to connecting the interaction center with the respective interaction surface point and doing so *within the same molecule* (see Fig. 2). In this way, we obtain all possible translation vectors by considering just a single molecule. The whole crystal must be spanned by three of these vectors (**a**, **b**, **c**), provided that it is closely packed and all relevant molecular interactions are represented by the model. Most naively, one would look for a crystal structure by repeatedly taking three vectors out of the candidate set and calculating the crystal energy. To render this approach computationally feasible, the search for the correct triple of vectors has to be structured. Towards this end, we impose several constraints on the triples we take into consideration.

(i) First, we strongly reduce the number of different vectors by using a clustering method. This step introduces certain inaccuracies into the process but, at the same time, it reduces the size of the search space sufficiently to render the subsequent search feasible.

There are several methods of adaptive clustering available (see, for instance, Rarey, Wefing & Lengauer, 1996). For the present application, we prefer a uniform clustering method that maps the vectors onto a three-dimensional mesh (*grid constraint*). The reasons for this choice are, first, that a mesh allows a natural procedure of estimating the accuracy of the model and, second, that the subsequent constraints, especially the triangle constraint, depend on a uniform discretization. We choose 1 Å as the mesh size. This value presents a reasonable trade off between accuracy and run time.

(ii) Second, we restrict our attention to vectors whose energy is small (*energy constraint for vectors*). For this purpose, the energy of the vectors is evaluated, the vectors are ranked with respect to their energy and only highest-ranking vectors, *i.e.* those with lowest energy, are taken into account. In our calculations, we retained the 500 best vectors.

(iii) Third, we require that at least two (**a**, **b** without loss of generality) of the three vectors have to fulfill the following constraint (*triangle constraint*): there must be a vector **d** among the 500 best ranking candidates such that

$$\mathbf{d} = \pm(\mathbf{a} - \mathbf{b}). \quad (1)$$

Vector **a** defines a one-dimensional chain. Vectors **b** and **d** represent interaction vectors of two molecules of the chain with one molecule in an adjacent chain. The vectors **a** and **b** define a crystal sheet. The relationship described by (1) is always fulfilled if the surfaces are strictly convex, the three-dimensional shapes are periodically close packed in a sheet, and any surface-to-surface vector of the shape can be selected. In Fig. 3, the situation is illustrated for discs, and Fig. 4 shows a counterexample for shapes that are not strictly convex.

(iv) Fourth, in further calculations, we consider only triangles whose energy is small (*energy constraint for triangles*). Analogously to the energy constraint for the vectors, the energy of the triangles, which is just the sum of the energy of the three vectors **a**, **b** and **d**, is calculated. The triangles are ranked with respect to their energy and the 200 highest ranking of them are retained. Each crystal structure is then built up from one of these two-dimensional triangles plus one of the 500 vectors, thereby completing the the three-dimensional unit cell.

(v) Fifth, we require the density to be within a certain range (*density constraint*). The density of most known organic crystals is in the range 1.1 to 1.6 g cm^{-3}. These are the values chosen in our calculations. An advantage of this constraint is that the density is an easily available property of crystals. If the density is known but the structure is unknown, *FlexCryst* can be used to calculate possible structures with the required density.

(vi) Sixth, we apply an *energy constraint for crystals*. The algorithm generates a large number of crystal structures by combining the vectors **a** and **b** satisfying the triangle constraint with a vector **c** satisfying the density constraint. The vectors **a** and **b** span the crystal plane and the vector **c** completes the elementary cell. The scoring function is evaluated for all crystal structures meeting all constraints. Afterwards, the crystal structures are ranked with respect to their energy. For the clustering process, we retain a certain number of these crystals – 1000 for space group $P1$ and 10 000 for $P\bar{1}$. The necessarily larger number of structures to be retained for $P\bar{1}$ is due to the larger number of degrees of freedom.

Finally, the energetically lowest structure of each cluster is compared to the experimental structure.

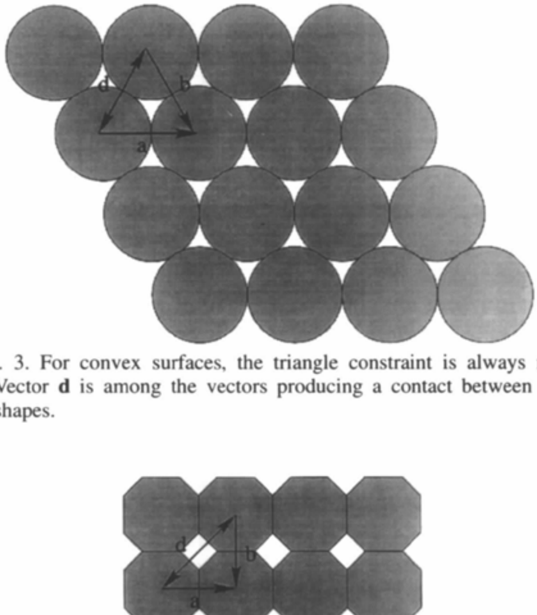


Fig. 3. For convex surfaces, the triangle constraint is always met. Vector **d** is among the vectors producing a contact between two shapes.

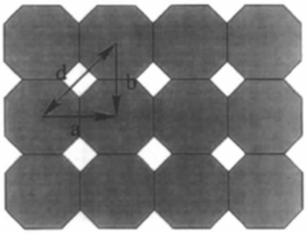


Fig. 4. For concave surfaces, the triangle constraint may be violated. Vector **d** is not among the vectors producing a contact between two shapes.

3. Scoring function

A well known deductive approach to protein structure prediction is to derive potentials from data on resolved molecular structures in a statistical fashion (Sippl, 1990, 1993; Sun, 1993). The similarity of Boltzmann statistics to the statistics of interactions was first noted for the distribution of the conformations of the residues of proteins over the angles φ, ψ and χ (Pohl, 1971). The theoretical background of this fact was formulated 20 years later (Gutin, Badretinov & Finkelstein, 1992). The idea of statistically derived potentials can be carried over to calculating the intermolecular potentials of organic compounds in crystals.

With the inverse Boltzmann equation, the potential energy between two interacting atoms A_{iI} and A_{jJ} of different molecules I and J (only intermolecular interactions are of interest here) can be written as

$$E_{\text{atom}}(r_0, i, I, j, J) = E_{ij} + N_L kT \log \lim_{r_\infty \to \infty} [P_{ij}(r_\infty) r_0^2 / P_{ij}(r_0) r_\infty^2] \qquad (2)$$

with

$$r_0 = |\mathbf{r}(A_{iI}) - \mathbf{r}(A_{jJ})|. \qquad (3)$$

$P_{ij}(r_0)$ is the probability that the shell at distance r_0 around an atom of type i contains an atom of type j and *vice versa*. $P_{ij}(r_\infty)$ is the probability of finding two atoms independently of each other, as in the case of an infinite distance between the two atoms. This probability can also be expressed by the average densities ρ_i and ρ_j of the atom types in the crystals.

$$\lim_{r_\infty \to \infty} [P(r_\infty)/r_\infty^2] \propto \rho_i * \rho_j. \qquad (4)$$

We estimated the value of the integration constant E_{ij} and the decoupled probability $P_{ij}(r_\infty)$ by the following procedure. We statistically derive the pair potential function with undetermined shift E_{ij}, applying (2) to the atom-pair correlation function. In order to have enough data to evaluate the atom-pair correlation function (as an example we show the function for H—H in Fig. 5), we used the Cambridge Structural Database (Allen & Kennard, 1993). We parameterized the most relevant interactions (see Fig. 6) and disregarded the contributions of other interactions. An extension to other chemical elements by providing the additional pair correlation functions of these elements is straightforward. The only limitation is the sparsity of available data for several interaction pairs. For each interaction, we evaluated the alphabetically first 1000 different crystals containing the corresponding interaction. This number of structures is sufficiently large for the calibration, as can be argued from the fact that the pair potential functions become almost constant for distances above 4.0 Å, see *e.g.* Fig. 9. This is to be expected for decoupled probabilities. For this reason, we replace $P(r_\infty)$ by the value of $P(4.0\,\text{Å})$ and disregard energy contributions for atom pairs with larger distances than 4.0 Å.

To determine E_{ij}, we made use of the fact that the volume of predicted crystals depends on E_{ij}. For increasing E_{ij}, the volume of the predicted crystals increases as well. This is caused by the mostly monotonically declining pair energy functions in the range of the van der Waals contacts. Calibration of an average shift E_{ij} for all pair interactions such that the predicted and experimental volumes of crystals considered are equal gives us a reasonable value for E_{ij}. For our training set, we obtain a value of $-2.85\,\text{kJ mole}^{-1}$. In Fig. 7, we plot the experimental *versus* the calculated volume. The straight line is the line of regression for the calculated volumes and should be identical to the dashed line on which the experimental volume is the same as the

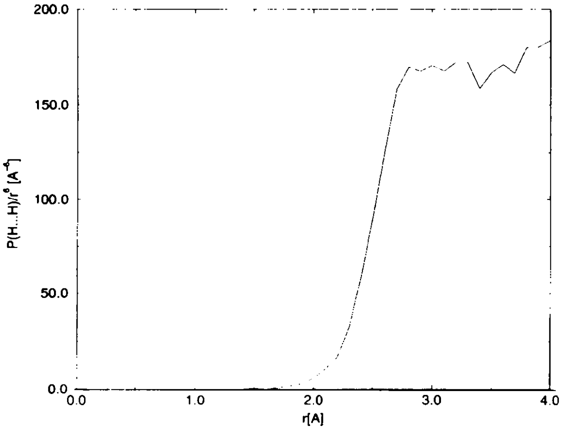


Fig. 5. The atom-pair correlation function for the H—H interaction. $P(\text{H}\cdots\text{H})$ is the probability of finding an intermolecular distance of r Å between two H atoms.

	H	C	N	O	F	P	S	Cl
H	◆	◆	◆	◆	◆	◆	◆	◆
C	◆	◆	◆	◆	◆	◆	◆	◆
N	◆	◆	◆	◆	◆		◆	
O	◆	◆	◆	◆	◆		◆	◆
F	◆	◆	◆	◆	◆			
P	◆	◆						
S	◆	◆	◆	◆			◆	
Cl	◆	◆		◆				◆

Fig. 6. A list of included atom-pair interactions. Parameterized atom pairs are indicated by ◆.

calculated volume. The influence of the value E_{ij} on the calculated volumes can be seen in Fig. 8, where a value of 3.35 kJ mole^{-1} was assumed for E_{ij}.

Replacing E_{ij} and $P(r_\infty)$, one can rewrite the inverse Boltzmann equation as

$$E_{\text{atom}}(r_0, i, I, j, J) = \begin{cases} -0.68\ \text{kcal mole}^{-1} + N_L kT \log \dfrac{P(4.0\ \text{Å}) r_0^2}{P(r_0) 4.0\ \text{Å}^2} & \text{if } r_0 \leq 4\ \text{Å} \\ 0\ \text{kcal mole}^{-1} & \text{if } r_0 > 4\ \text{Å}. \end{cases} \quad (5)$$

As one example of the determined pair energy functions, we show the potential function for H—H in Fig. 9. This curve deviates considerably from classical pair potential functions; in particular, one cannot recognize a clear van der Waals minimum. However, note that this function includes the interatomic stabilization of the surface interactions and, in addition, the molecules' interior intramolecular stabilization, which results in the unexpected shape of the curve.

The hydrogen bond provides an example of an interaction that involves pairs of atomic groups rather than just pairs of single atoms. The hydrogen bond can be modeled cooperatively by the atom-pair potentials for H—O and O—O, see Fig. 10. In hydrogen bonds, one H—O distance is around 0.9 Å, the other is around 2.0 Å and the H—O—H angle is around 180°. The H—O potential has a minimum around 2.0 Å and the O—O potential has a minimum at 2.8 Å. These minima model the geometry of the hydrogen bond quite accurately, *i.e.* to within about 0.1 Å.

The pair potential of O—O has a high-energy local minimum at 2.0 Å. This minimum arises from a single observed O—O distance of 2.0 Å in the Cambridge

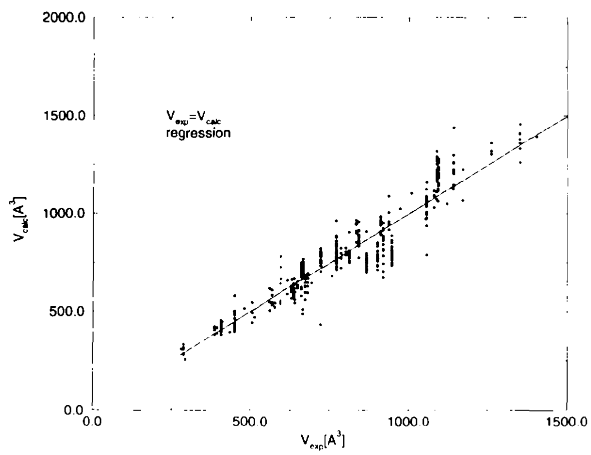


Fig. 7. Comparison of calculated and experimental volumes for E_{ij} = 2.85 kJ mole^{-1}.

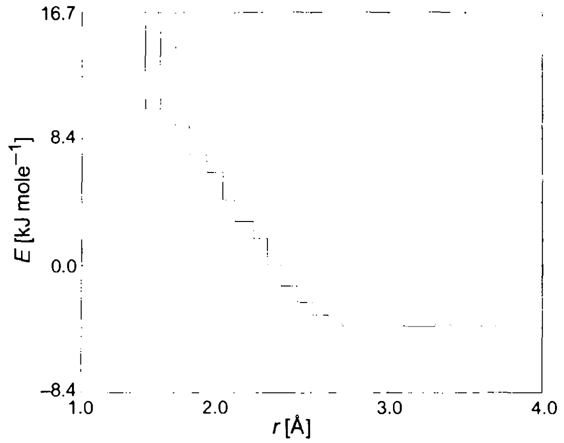


Fig. 9. The atom-pair potential function for the H—H interaction, obtained by applying the inverse Boltzmann equation to the atom-pair correlation shown in Fig. 5.

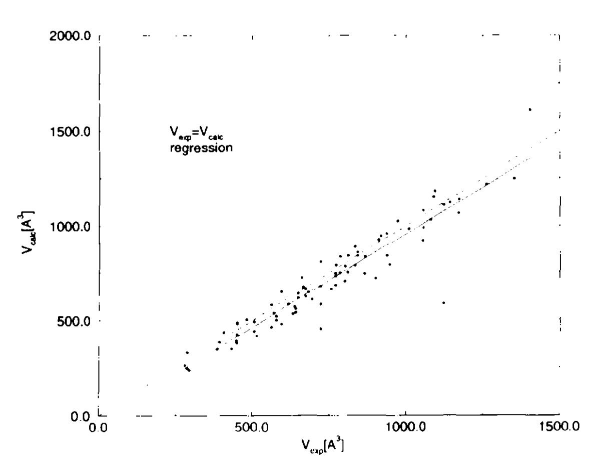


Fig. 8. Comparison of calculated and experimental volumes for E_{ij} = 3.35 kJ mole^{-1}.

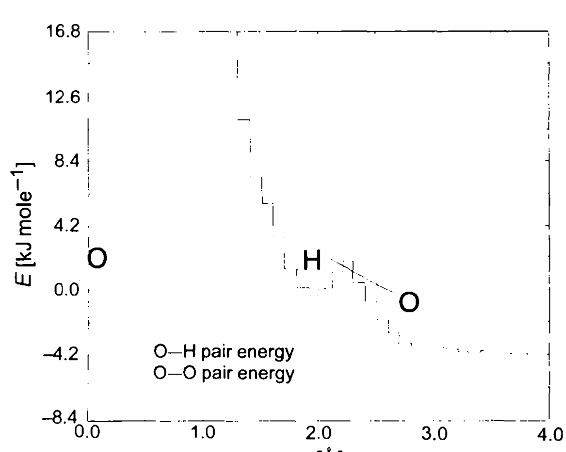


Fig. 10. Modeling hydrogen bonds with the O—O and O—H atom-pair potential functions.

Structural Database. We assume that the corresponding structure is in error. Such problems occur with other types of interaction as well. However, such minima influence the potential slightly because their energy is large.

The intermolecular interaction energy between two molecules I and J with n_0 atoms is formulated as a sum of atom-pair interactions:

$$E_{\text{dimer}}(I, J) = \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} E_{\text{atom}}(i, I, j, J). \qquad (6)$$

The total stabilization energy of one molecule in the crystal is formulated as a sum of dimer energies:

$$E_{\text{crystal}}(I) = \sum_{J=1}^{N} E_{\text{dimer}}(I, J). \qquad (7)$$

The number N of pairs to consider depends on the number of adjacent cells n_{neighbor} in each of the three directions to be included. N and n_{neighbor} are related as follows:

$$N = [(2n_{\text{neighbor}} + 1)^3 - 1]/2. \qquad (8)$$

In our calculations, only first neighbors have been considered. Therefore, $N = 13$. Though the elementary cell is surrounded by 26 cells, only half of the interactions must be taken into account because the others are related by symmetry.

4. Similarity and clustering of crystals

A widely discussed problem in the literature is how to define the notion of similarity between different crystal structures. With respect to our problem, we can limit ourselves to crystals formed by a single molecule. At present, two approaches are most widely used. The first approach (Karfunkel, Rohde, Leusen, Gdanitz & Rihs, 1993) takes advantage of existing programs for comparison between spectra. A spectrum transforms direct space, containing the coordinates of the atoms, to frequency space, describing the spatial periodicity of the atoms. The second approach is to present the unit cell in a normal form of the unit cell (Parthé & Gelato, 1984) and to calculate the square deviation between the atoms of the two unit cells (Burzlaff & Rothammel, 1992; Dzyabchenko, 1994).

For our purposes, we propose a third method, which exploits the fact that we are always dealing with the same molecule, which is rigid and fixed in space. Our algorithm constructs the unit cell of the crystal of just this reference molecule. For the comparison of two structures, the problem is reduced to comparison of the corresponding unit cell. In the case of $P1$, this means that the translation vectors must be similar. For the space group $P\bar{1}$, the inversion center also has to be taken into consideration. We first concentrate on how to check the identity for the space group $P1$.

Let as assume that two crystal structures of the same molecule are described by the bases $\mathbf{B}$ and $\mathbf{B}'$, respectively. Each of these bases consists of three translation vectors:

$$\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3), \quad \mathbf{B}' = (\mathbf{b}'_1, \mathbf{b}'_2, \mathbf{b}'_3). \qquad (9)$$

Each of these sets of vectors define a point lattice $\mathbf{X}$:

$$\mathbf{X} = \mathbf{Bn} : \mathbf{n} \in \mathbb{Z}^3. \qquad (10)$$

A lattice X' is a sublattice of a lattice X if, for $i = 1, 2, 3$, $\mathbf{b}'_i$ is an integer linear combination of the vectors $\mathbf{b}_j$:

$$\mathbf{X}' \subseteq \mathbf{X} \quad \text{if} \quad \exists \mathbf{t}_i : \mathbf{Bt}_i = \mathbf{b}'_i \wedge \mathbf{t}_i \in \mathbb{Z}^3 \wedge i \in \{1, 2, 3\}. \qquad (11)$$

This condition is easy to check. To prove the identity of the two lattices, it would be necessary to check this condition in both directions. Proper subsets, necessarily describing low-density crystals, are already screened out from the possible crystal structures by the *density constraint*, however.

To define a similarity index, we introduce the quantity s:

$$s = \max\{|\mathbf{Br}_i|\} \quad \text{with} \quad r_{ij} = \lfloor t_{ij} + 0.5 \rfloor - t_{ij}, \quad i \in \{1, 2, 3\}, \quad j \in \{1, 2, 3\}. \qquad (12)$$

That this quantity is an useful notion of similarity can be seen as follows. If $X = X'$, s is zero. If we change continuously one of the translation vectors of lattice X', the function increases monotonically. In our calculations, we consider crystals to be similar as long as s is below a certain threshold (1 Å). This coincides with the mesh size for discretization. As a consequence, crystals are considered to be similar if each translation vector of X is identical or adjacent to a translation vector of X' on the mesh of discretization (diagonals excluded).

For the case of $P\bar{1}$, one more vector has to be added on the right-hand side of the similarity quantity s. Two inversion centers are symmetry related and will not be distinguished if twice the difference of their vectors $\mathbf{b}_4$ and $\mathbf{b}'_4$ can be expressed as an integer linear combination t_4 of the basis:

$$2(\mathbf{b}_4 - \mathbf{b}'_4) = \mathbf{Bt}_4. \qquad (13)$$

Analogously to (12), r_4 is defined and the quantity s extended.

$$s = \max\{|\mathbf{Br}_i|\} \quad \text{with} \quad r_{ij} = \lfloor t_{ij} + 0.5 \rfloor - t_{ij}, \quad i \in \{1, 2, 3, 4\}, \quad j \in \{1, 2, 3\}. \qquad (14)$$

5. Results for *P*1

To validate the algorithm, we extracted a set of 131 crystals of space group *P*1 from the Cambridge Structural Database. We selected all organic crystals containing only the elements H, C, N, O, F, P, S and Cl of space group *P*1. The complete search expression can be found in Appendix *A*. The molecular structure formulas were given as input to the program and the generated candidates for the three-dimensional crystal structure were compared to the experimentally observed crystal structure. As is well known, H atoms are often missing or are not precisely located in crystallographic data. For this reason, the data were read out from the Cambridge Structural Database and the *SYBYL* system (TRIPOS Associates, 1994) was used to add missing H atoms. Then we input the hydrogen-completed structures to *FlexX*. The interaction centers and interaction surfaces were calculated with *FlexX*. These results were collected in an interaction file containing the interaction surface points and interaction centers. The program *FlexCryst* calculated from that interaction file and the molecular information of the structure file the various potential crystal structures. Finally, the crystallographic data, cell length and cell angles of the structure file were used to compare the computed structures to the observed crystal structure.

In 129 of 131 cases (98%) (Fig. 11), the program succeeded in finding a structure that is either identical or similar to the experimental structure. Here we applied

crystal structure was detected by FlexCryst (129 cases=98%)				
ADGSMF(1)	ADGSMH(1)	BADVOD10(1)	BAKHOK(2)	BDORLA10(10)
BEKHUU(19)	BERVEZ(1)	BETJEP(1)	BIPPEV(4)	BIXHOF(1)
BOTSAE(1)	BXCPAF(2)	CEGLCA(6)	CEGLCA01(2)	CERPAQ(31)
CETROI(3)	CETROI01(8)	CIFYOF(1)	CIFYOF10(1)	CIYRIL01(1)
COMCIQ(1)	COTCIX(1)	CUVFOO(1)	DAKSAJ(2)	DARNUF(1)
DEBLOL(1)	DERCIM(3)	DIGOXN(7)	DIGOXN10(4)	DIWXIQ(1)
DOHHIR(6)	DOZMIO(1)	DUMCET(1)	EACJEX(1)	ECPRPR01(16)
FADGEW(1)	FAKGAZ01(1)	FALKAE(16)	FAMDUS(4)	FATXUT(17)
FAVSUQ(1)	FEPZOP(4)	FETWOQ(6)	FEXCOA(1)	FITVOT(1)
FIYJIG(1)	FOMANN(1)	FUNVUF(4)	FUPVAN(3)	FUXBIJ(93)
FUXBIJ01(50)	GEYMEC(1)	GIPJEU(33)	GOJHIW(3)	HAGFAW(2)
HCARDO(1)	HCARDO01(1)	HOLOTM(8)	HPICRB(2)	HTENTX10(1)
JANDUX(1)	JECYIZ(1)	JIHREX(1)	JIJXEF(1)	JIPBIT(1)
JOVZAV(1)	JUFTUZ(1)	KANDUY(1)	KANTOI(7)	KEGBAZ(5)
KERSIJ(1)	KIJCAH(3)	KITLUU(1)	KOCHIT(1)	KOHNAW(1)
KOPROW(12)	LAWKUP(9)	LCDMPP01(13)	LCDMPP10(29)	LEDNUD(2)
LEKVIG(1)	LEMZAE(12)	LETBOB(1)	LYSDOL(4)	MAMNAC(1)
NALCYS02(1)	OACGAP(1)	OHWTHN(1)	OMAPBD(1)	PAJSOI(1)
PATCUI(1)	PATPYS(1)	PEVLOR(3)	PICSEZ(124)	PICTIE(5)
PIKYIR(14)	PMNTBZ(2)	PROGLE20(2)	RPPYPY20(3)	SESHUT11(2)
SEZLUE(4)	TEOXDE01(8)	THPGFA(2)	VARHUR(6)	VARWUG(4)
VEGJOG(1)	VEKZAM(1)	VITREV(4)	VOBHEZ(1)	VOBPEH(1)
VOBPEH10(1)	VOFFAX(131)	VOXXUB(1)	VOYVEK(1)	WATCID(47)
WICVUZ(4)	WIKSEO(22)	WINWEV(1)	YABVUS(1)	YAMBET(1)
YEBGIV(17)	YEHRIM(1)	YIJBUO(1)	YIPPAO(3)	YIPWAV(1)
YOGVOF(1)	YOKGIO(1)	YUYHAB(9)	ZAYWIJ(1)	
crystal structure was not detected by FlexCryst (2 cases=2%)				
FURCOU	CILWOJ			

Fig. 11. A complete list of the reference codes of calculated crystals of space group *P*1. If a structure was found that is similar to the experimentally observed structure, the rank is given in parentheses.

the notion of similarity described in §4. To check the reliability of the scoring function, the ranks of computed structures that were similar to the experimentally observed structure were calculated. Fig. 12 shows a plot of the result. In 68 cases, the experimentally observed structure is similar (in the above sense) to the minimally scoring computed structure. In 59 cases, there is a structure among the computed structures with ranks between 2 and 100 that is similar to the experimentally observed structure. In two cases, the highest-ranking similar computed structure has a rank above 100. Only in two cases did the program not find the experimental structure. In one of these cases (CILWOJ), the structure might be in error (Bocelli, 1986). Our calculations confirm this suggestion. The reason for the other case is still under investigation.*

A complete list of the reference codes of the calculated crystals is provided in Fig 11. The rank of the crystals found is given in parentheses.

For all 131 crystals together, the elapsed computing time was 40 min (average time per crystal = 18 s) on a SUN^{TM}Ultra^{TM}1 Workstation. The main reasons for the high speed compared to existing algorithms are that we avoid repeated optimization and that we use statistically calibrated potentials. It is not necessary to evaluate terms like Ar^{-6} and Br^{-12} (Lennard-Jones potential) or Ar^{-6} and Be^{-Cr} (Buckingham potential), respectively, in the most critical step.

6. Results for P$\bar{1}$

Our test set for space group $P\bar{1}$ consists of the first 95 entries of the CSD, with the same limitations as before. In particular, this means that the compound consists only of the elements H, C, N, O, F, P, S and Cl. We applied the same options as in Appendix *A* for space group $P1$ with the small differences listed there.

The program succeeded in 81 cases (85%). In 11 cases, the crystal with the lowest rank is similar to the experimental structure. For 76 cases, the experimental structure is among the first 1000 candidates. In five cases, the highest-ranking similar computed structure has a rank above 1000 up to 10 000. In 14 cases, the program did not find the experimental structure. The reasons for the latter set are still under investigation. In some cases, reasons include missing parametrized groups in *FlexX* (—CN in ACAMEL), lack of parametrized atom-pair potentials, as is the case for BABHAN containing deuterium. Sometimes, one of the constraints needed to be relaxed. For example, the compound ATZCXB has a density of $\rho = 1.89\,\mathrm{g\,cm^{-3}}$, which is outside our allowed range. Very lengthy molecules (ALAEUC10) are violating the *energy constraint for vectors* because vectors connecting two chain ends are much higher in energy than vectors representing side-to-side contacts and these high-energy vectors are excluded by the energy constraint in the algorithm (Fig. 13).

The ranking of the lowest-energy structure found to be similar to the experimental structure shows a much more diffuse pattern (Fig. 12) than in the case of $P1$.

A full list of all considered entries of the CSD is given in Fig. 14.

The larger number of degrees of freedom in the group $P\bar{1}$ causes longer run times for crystal structures of this space group than for the simpler space group $P1$. The overall time was 136 h, which gives an average run time of 86 min per crystal. These results are still much faster and/or more accurate than other algorithms known to

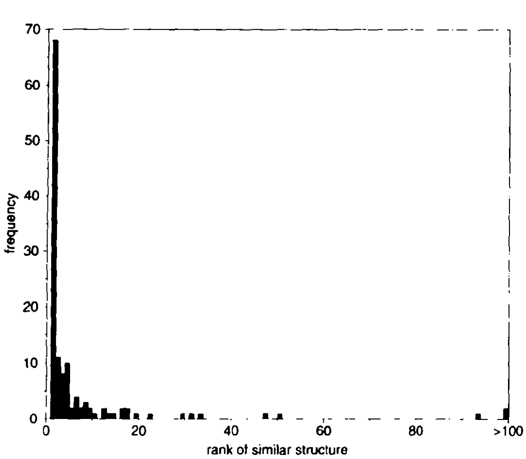


Fig. 12. The lowest rank of any computed crystal structure that is similar to the experimentally observed crystal structure. For two crystals, this rank is larger than 100.

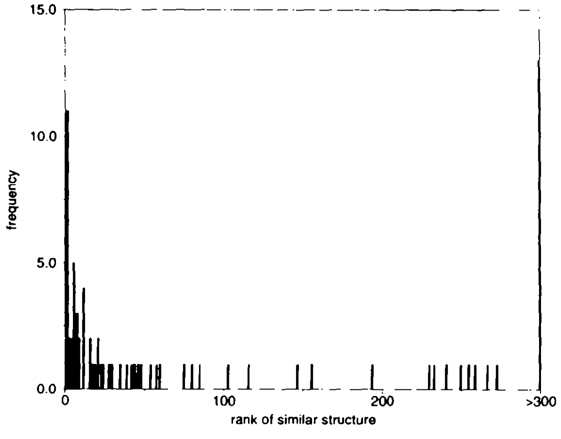


Fig. 13. The lowest rank of any computed crystal structure that is similar to the experimentally observed crystal structure. For 12 crystals, this rank is larger than 300.

* In the course of our tests, we originally found 49 structures in the CSD that erroneously were reported to belong to space group $P1$ instead of $P\bar{1}$ or contained metals and, nevertheless, were flagged to be organics instead of metal-organics. In general, we reported such errors to the CSD group immediately and corrections in the databank were performed quickly.

the authors. A recently published article favorably rates a novel method by Chaka, Zaniewski, Youngs, Tessier & Klopmann (1996) to a set of other methods. The computation time for the best method given in that paper is 2 h per structure on an CRAY-YMP for a parallelized version of the program. The probability of finding the crystal structure is reported to be 10 out of 16 (62%) for a manually selected set of crystals.

7. Conclusions

We have presented a discrete algorithm that detects the experimentally observed structure among the computed candidates in almost every case for the simple case of $P1$ with $Z = 1$ and in a large percentage of the cases of $P\bar{1}$ with $Z = 2$.

The speed of the program is several orders of magnitude faster than existing algorithms. Three ingredients are essential for this efficiency: performing an analysis of the intermolecular interaction as a preprocessing step, using a discrete configuration space, and reducing the energy calculation to simple table look-up.

As the potentials are not restricted to predefined classes of functions, pair potentials with unusual forms are also possible, an example being the O—O pair potential with two minima.

Instead of randomly choosing crystals for the statistics as we did, the potential can be adapted to a certain class of substances, *e.g.* only aromatic compounds, by including only crystals of the respective class into the statistical analysis.

Replacing the atoms by functional groups is an easy way to extend the potentials from two-body interactions (atom–atom) to many-body interactions.

The authors thank Dr P. Erk (BASF) for fruitful discussions and Dr H. Slot (University of Nijmegen) for kindly assisting them in using the Cambridge Structural Database. Furthermore, we are grateful to P. Lauwers, C. Lemmen, B. Kramer, C. Oligschläger, M. Rarey and S. Wefing for helpful comments on the paper. This research was performed as part of GMD's contribution to HLRZ (High Performance Computing Center).

crystal structure was detected by FlexCryst (81 cases=85%)				
AABHTZ(406)	ABZNPS(59)	ACAMOX(512)	ACINDD(57)	ACMGHX(11)
ACNODC(1)	ACNPTH(74)	ACNPTH10(11)	ACOHKT(4)	ACPSTP(1)
ACSLCB10(240)	ACTCBZ(34)	ACTHBZ(229)	ACVCHO(2)	ADYPNL(8)
AISMRS(23)	ALXANM01(511)	ALXANM10(29)	AMFPDO10(27)	AMHPEN10(5)
AMIMZA10(1503)	AMXBPM10(1976)	AMYTZL(16)	ANTSTK(15)	AOPCHY(1)
APBTRN(7)	APTSPN(5)	ATBXYL10(5)	ATDZSA(387)	ATZCXB(3127)
AXBCHX(524)	AZNAND(1)	BAFKUO(43)	BAFLEZ(193)	BAGKID10(47)
BAGPOO(372)	BAGWUB(18)	BAGXUC(53)	BAHFIZ(115)	BAHYUE(155)
BAJRUZ10(7)	BAJWIS(3)	BAKCOF(1086)	BAKDIA(84)	BAKXOA(5)
BAMFOK(15)	BAMFUQ(146)	BANGAY(11)	BANLUW(38)	BARCAY(22)
BARGOQ(1)	BASVOG(1)	BASXIC(309)	BASXIC10(232)	BAWMIV(720)
BAYHIS(5)	BAYNEU(1)	BAZBIN(102)	BDTPIM(11)	BEBCEQ(2)
BEBCUG(20)	BEBLOJ(7)	BEBMAW(254)	BEBWOU(249)	BEBWUA(1986)
BEBYEM(272)	BECJEY(6)	BECWUB10(17)	BEDFAR(258)	BEDHAT(45)
BEDKIE(4)	BEDXOX(1)	BEFSIO(266)	BEGBAQ(41)	BEGCIZ(1)
BEHCAS(8)	BEHMHD(20)	BEHXOB(79)	BEJJIJ(1)	BEJKEG(1)
BEJMIM(6)				

crystal structure was not detected by FlexCryst				
reason unclear (14 cases=15%)				
ACAMEL	ACMCPY	ACYACR	ALEUAC10	AMPCTZ
APFPTS	BABHAN	BABROL	BAGXAI	BALRIP
BARZUP	BAZVON	BDTPAE	BEGDEW	

Fig. 14. A complete list of the reference codes of calculated crystals of space group $P\bar{1}$. If a structure was found that is similar to the experimentally observed structure, the rank is given in parentheses.

APPENDIX *A*

The complete search expression for extracting structures of space group *P*1 is given. We used *QUEST3D*.

```
T1 *CONN
NFRAG   1
AT1 H 0                                        :XY   496   496
END
T2 *CONN
NFRAG   1
AT1 Br 0                                       :XY   496   496
END
T3 *CONN
NFRAG   1
AT1 B 0                                        :XY   496   496
END
T4 *CONN
NFRAG   1
AT1 As 0                                       :XY   496   496
END
T5 *CONN
NFRAG   1
AT1 I 0                                        :XY   496   496
END
T6 *CONN
NFRAG   1
AT1 Si 0                                       :XY   496   496
END
T7 *CONN
NFRAG   1
AT1 Se 0                                       :XY   496   496
END
COMMENT Turning ON "INSIST-ON-COORDS"
SCREEN  153
COMMENT Turning ON "INSIST-NO-POLYMERS"
SCREEN  -54
COMMENT Turning ON "INSIST-NO-DISORDER"
SCREEN   35
COMMENT Turning ON "INSIST-PERFECT-MATCH"
SCREEN   85
COMMENT Turning ON "INSIST-RFACTOR<=10%"
SCREEN   88
COMMENT Turning ON "INSIST-ERROR-FREE"
SCREEN   33
SCREEN   57
SAVE FDAT
T8 *SPACEGROUP 'p1 '
T9 *ZVALUE .EQ. 1.0000
T10 *NRESIDUES .EQ. 1
QUEST
T8.AND.T9.AND.T10.AND.T1.NOT.T2.NOT.T3.NOT.T4.NOT.T5.NOT.T6.NOT.T7
```

For space group $P\bar{1}$, ZVALUE.EQ.2 and in addition SCREEN -154 was used.

References

Allen, F. H. & Kennard, O. (1993). *Chem. Des. Autom. News*, **8**, 1, 31–37.

Bocelli, G. (1986). *Acta Cryst.* C**42**, 127–128.

Böhm, H.-J. (1992). *J. Comput. Aided Mol. Des.* **6**, 593–606.

Böhm, H.-J. (1994). *J. Comput. Aided Mol. Des.* **8**, 243–256.

Borchardt-Ott, W. (1976). Editor. *Kristallographie, eine Einführung für Naturwissenschaftler*, pp. 89–95. Berlin: Springer-Verlag.

Burzlaff, H. & Rothammel, W. (1992). *Acta Cryst.* A**48**, 483–490.

Chaka, A. M., Zaniewski R., Youngs, W., Tessier, C. & Klopmann, G. (1996). *Acta Cryst.* B**52**, 165–183.

Desiraju, G. R. (1989). Editor. *Crystal Engineering, the Design of Organic Solids*. Amsterdam: Elsevier.

Dzyabchenko, A. V. (1994). *Acta Cryst.* B**50**, 414–425.

Eijck, B. P. van, Mooij, W. T. M. & Kroon, J. (1995). *Acta Cryst.* B**51**, 99–103.

Filippini, G. & Gavezotti, A. (1993). *Acta Cryst.* B**49**, 868–880.

Gutin, A. M., Badretinov, A. Y. & Finkelstein, A. V. (1992). *Mol. Biol.* **26**, 94–102.

Hahn, Th. & Klapper, H. (1995). *International Tables for Crystallography*, Vol. A, 4th ed., edited by Th. Hahn, Sections 10.5.5–10.5.6. Dordrecht: Kluwer.

Hill, T. L. (1948). *J. Chem. Phys.* **16**, 938–949.

Karfunkel, H. R., Rohde, B., Leusen, F. J. J., Gdanitz, R. J. & Rihs, G. (1993). *J. Comput. Chem.* **14**, 1125–1135.

Kitaigorodskii, A. I. (1951). *Izv. Akad. Nauk SSSR Ser. Fiz.* **15**, 157–163.

Klebe, G. & Mietzner, T. (1994). *J. Comput. Aided Mol. Des.* **8**, 583–606.

Parthé, E. & Gelato, L. M. (1984). *Acta Cryst.* A**40**, 169–183.

Pohl, M. M. (1971). *Nature (London) New Biol.* **234**, 277–279.

Rarey, M., Kramer, B. & Lengauer, T. (1995). *Third International Symposium on Intelligent Systems for Molecular Biology*, edited by C. Rawlings *et al.*, pp. 300–308. Menlo Park: AAAI Press.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). *J. Mol. Biol.* **261**, 470–489.

Rarey, M., Wefing, S. & Lengauer, T. (1996). *J. Comput. Aided Mol. Des.* **10**, 41–54.

Shoda, T., Yamahara, K., Okazaki, K. & Williams, D. E. (1994). *J. Mol. Struc. (Theochem.)*, **313**, 321–334.

Shoda, T., Yamahara, K., Okazaki, K. & Williams, D. E. (1995). *J. Mol. Struc. (Theochem.)*, **333**, 267–274.

Sippl, M. J. (1990). *J. Mol. Biol.* **213**, 859–883.

Sippl, M. J. (1993). *J. Comput. Aided Mol. Des.* **7**, 473–501.

Sun, S. (1993). *Protein Sci.* **2**, 762–785.

Tajima, N., Tanaka, T., Arikawa, T., Sakurai, T., Teramae, S. & Hirano, T. (1995). *Bull. Chem. Soc. Jpn*, **68**, 519–527.

TRIPOS Associates (1994). *SYBYL*. TRIPOS Associates, Inc., St. Louis, Missouri, USA.

Veciana, J., Cirujeda, J., Rovira, C. & Vidal-Goncedo, J. (1995). *Adv. Mater.* **7**, 221–225.

Westheimer, F. H. & Mayer, J. E. (1946). *J. Chem. Phys.* **14**, 733–738.

Williams (1983). Editor. *Non-linear Optical Properties of Organic and Polymeric Materials, ACS Symposium Series* No. 223. Washington: American Chemical Society.

Xiao, Y. & Williams, D. E. (1993). *Chem. Phys. Lett.* **215**, 17–24.

Zink, I. (1978). *Acc. Chem. Res.* **11**, 289–295.